

Measuring Individual and Group Collaboration Skills using LLM-Annotated Conversational Process Data

Beyond Results 2026 Workshop: Process Data and AI in Educational Assessment

Surina He, Assistant Professor
Nanyang Technological University, Singapore
24 September 2026

Background



Measurement

Collaboration and teamwork have a long research history.

Collaboration are mostly measured by outcomes (team performance), through observation or self-report.



CSCL

Computer-supported collaborative learning (CSCL) studies examine collaboration through process data, such as discourse and digital traces.

They focus on how people collaborate with each other or with agents, and how collaboration enhances learning.



The gap

Few studies have focused on using process information to measure collaboration skills. But collaboration is a process-oriented construct.

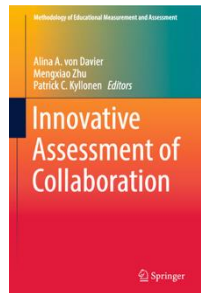
This study aims to measure collaboration skills using process data, such as collaboration discourse and turn-taking behaviors.

How?

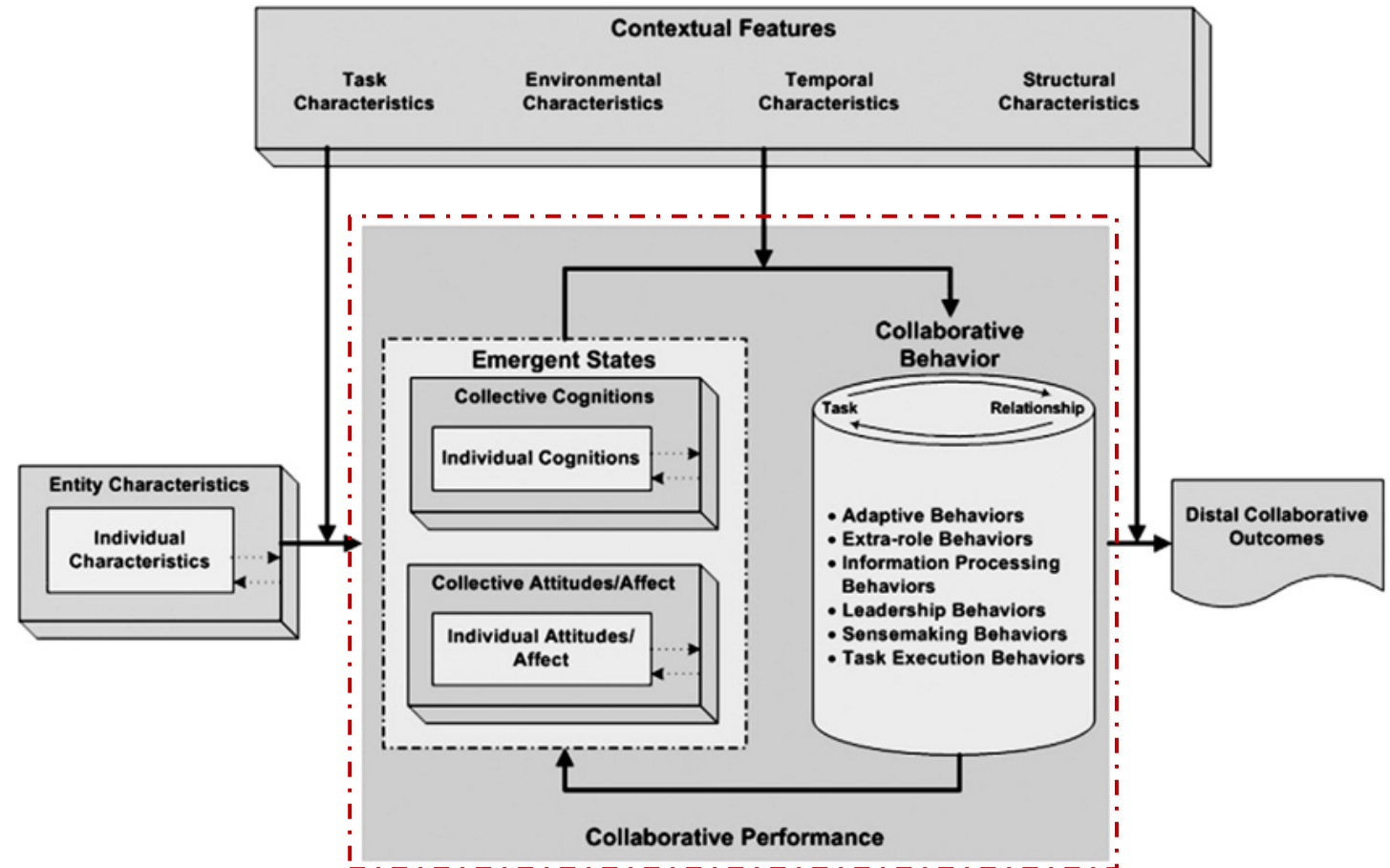
Background

Theories and multidisciplinary research on collaboration provide a solid theoretical grounding for building indicators from collaborative discourse.

Measurement of collaboration should capture the ABCs of collaboration: **Attitudes, Behaviors, and Cognitions.**

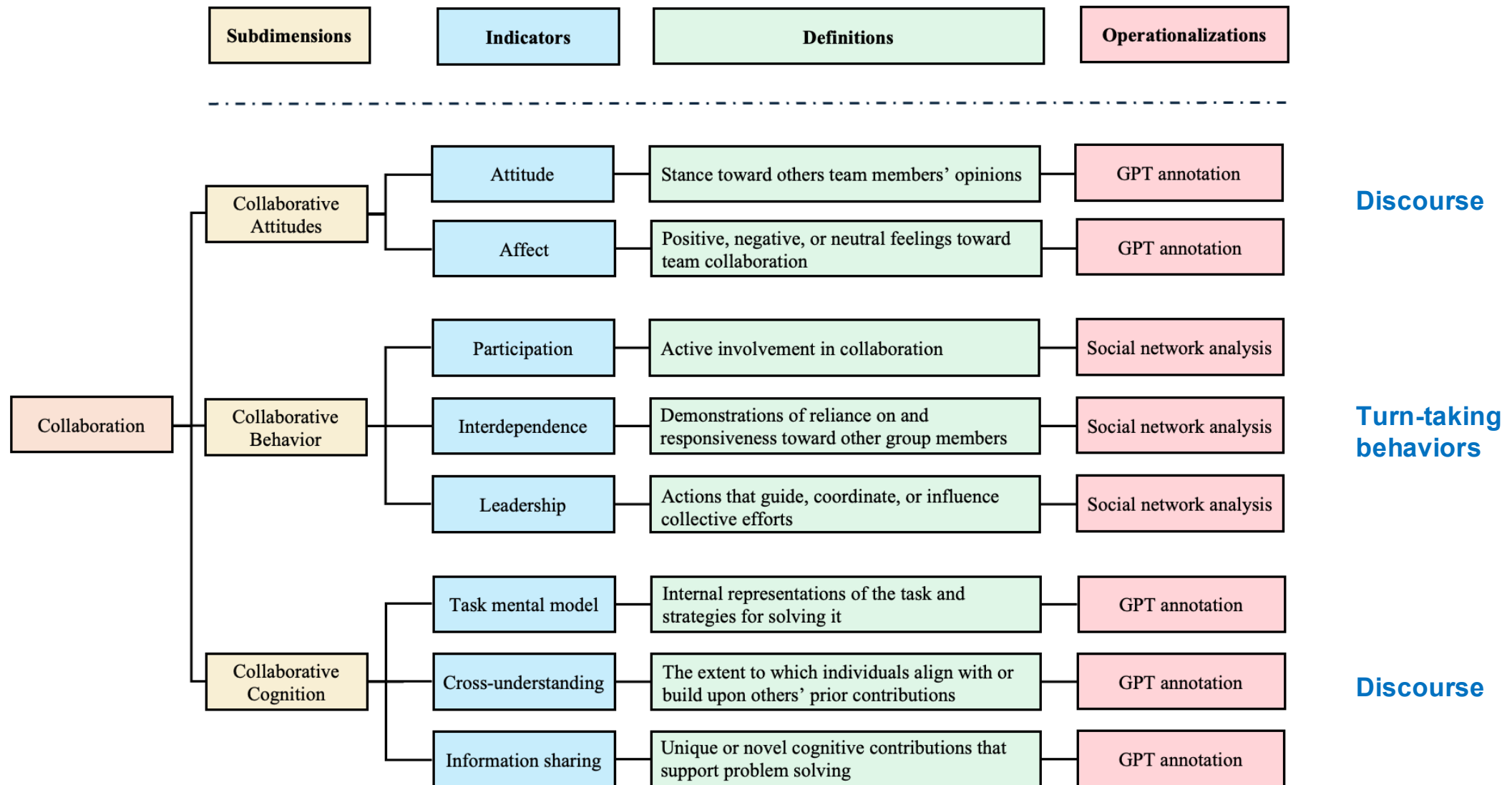


Innovative Assessment of Collaboration (von Davier et al., 2017)



Collaborative Performance Framework (Bedwell et al., 2012)

Method - Framework



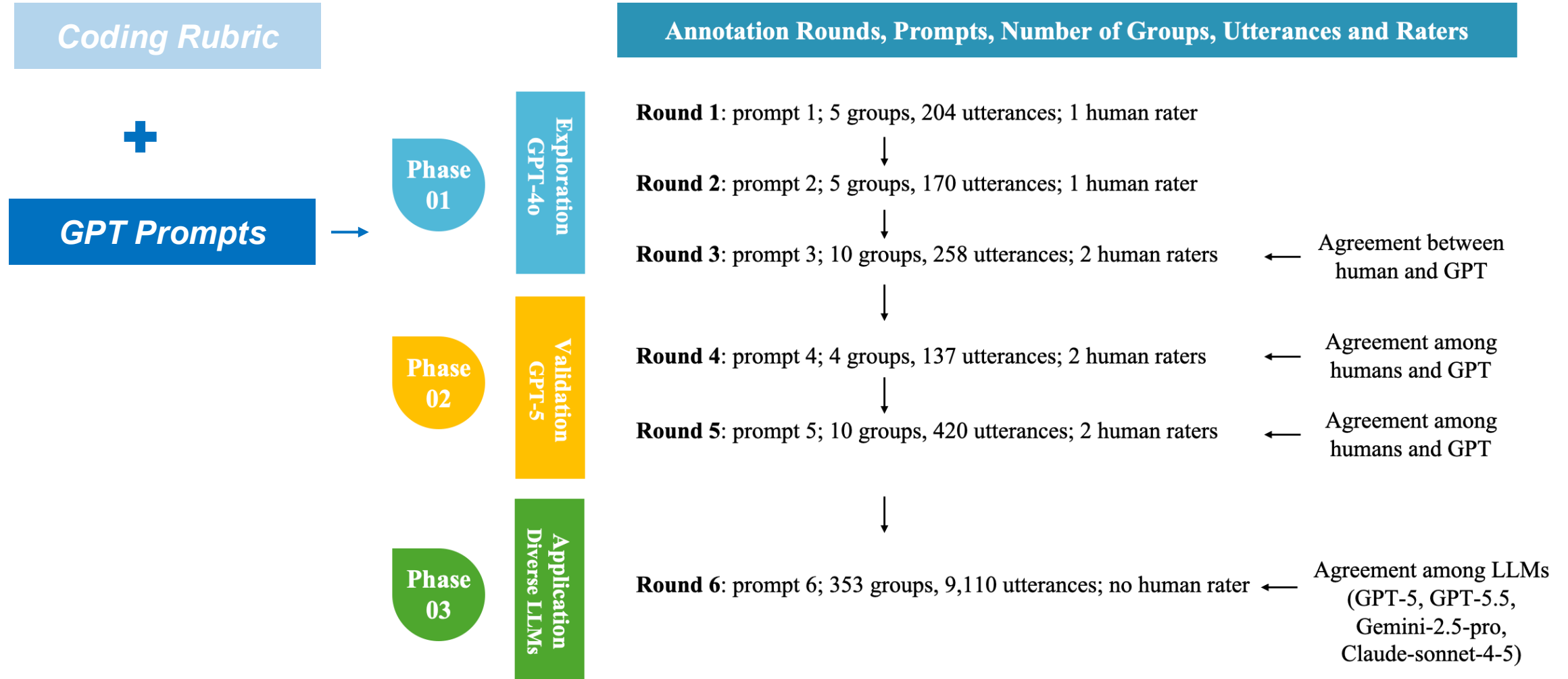
Method - Annotation

Coding Rubric



Indicator	Code	Label	Description
Affect	1	Negative	The utterance conveys negative affect, such as frustration, disappointment, or confusion.
	2	Neutral	The utterance expresses no emotional tone; it is factual, functional, or procedural in nature.
	3	Positive	The utterance conveys positive affect, such as encouragement, satisfaction, or enthusiasm.
Attitudes	1	Disagreement	The utterance explicitly or implicitly challenges, refutes, or contradicts previous team member's opinion.
	2	No Stance	The utterance expresses no clear agreement or disagreement; it may state a fact or describe an action.
	3	Agreement	The utterance explicitly expresses support, affirmation, or builds upon a previous contribution.
Information Sharing	1	Repetition	The utterance restates previously mentioned content without introducing new reasoning or ideas.
	2	Moderate Extension	The utterance slightly elaborates on existing ideas without introducing entirely new content. Restating task instructions for clarification also belongs to the category.
	3	New Contribution	The utterance introduces new reasoning, information, or perspectives to advance group understanding.
Cross Understanding	1	Mis-understanding	The utterance contradicts, overlooks, or misinterprets others' arguments, showing no integration with prior discussion, or shifts the focus to another topic.
	2	Partial Understanding	The utterance loosely connects to or vaguely references previous messages without clear synthesis. Brief expressions of agreement without elaboration also fit here.
	3	Full Understanding	The utterance clearly builds upon, extends, or synthesizes earlier contributions from others.
Task Mental Model	1	Off-Topic	The utterance is unrelated to the problem-solving objective and does not contribute to the task.
	2	Somewhat Relevant	The utterance mentions task-related elements but lacks a clear link to solving the problem.
	3	Clearly Relevant	The utterance directly contributes to solving the task, such as discussing instructions, strategies, and reasoning.

Method - Annotation



Method - Annotation

Coding Rubric



GPT Prompts



Annotated Output



GID	MID	Participant	Utterance	Annotation
G001	17	Cat	I am checking 7 because if it is a vowel, it disproves the rule.	• Affect: 2 Reason: Neutral action statement. • Attitudes: 2 Reason: Does not reference others' stances. • Information sharing: 1 Reason: Repeats earlier rationale about 7. • Cross understanding: 1 Reason: No explicit linkage to others. • Task mental model: 3 Reason: Directly applies the disconfirmation test.
G001	18	Kiwi	Ok You carry on.	• Affect: 3 Reason: Supportive encouragement. • Attitudes: 3 Reason: Supports Cat proceeding. • Information sharing: 1 Reason: No new reasoning. • Cross understanding: 2 Reason: Acknowledges prior action plan. • Task mental model: 2 Reason: Procedural encouragement, not reasoning.

Method - SNA



Log-transformed out-degree centrality

$$\text{Participation}_i = \log(1 + d_i^{\text{out}})$$

Normalized reciprocity

$$\text{Interdependence}_i = \frac{\sum_j m_{ij}}{d_i^{\text{out}}}$$

Normalized closeness centrality

$$\text{Leadership}_i = \frac{n - 1}{\sum_{j \neq i} d(i, j)}$$

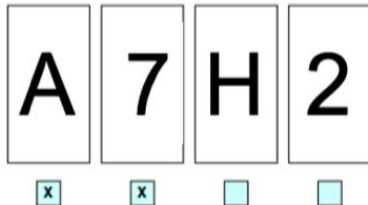
- d_i^{out} : Total number of outgoing transitions
- $d(i, j)$: Shortest-path distance between individuals i and j

Method - Data

- Deliberation Enhancing Bot project, led by researchers from University of Cambridge
- First publically available multi-person collaborative problem-solving data
- 1,637 adults participants in 353 groups (3-6 person per group) with over 9,000 utterances
- Wason card selection task:

Each of the 4 cards below has a letter on one side and a number on the other. Which card(s) do you need to turn to test the rule:

All cards with vowels on one side have an even number on the other.



- ① **Zebra:** What answers did everyone put?
- ② **Beaver:** I put A and 2
- ③ **Duck:** I put A and 2, but it may not be as simple as we think
- ④ **Beaver:** So is it 7 ?
- ⑤ **Zebra:** if we turn the 7 over & their isn't a vowel we test its true
- ⑥ **Duck:** I think 7 and A
- ⑦ **Beaver:** is it 2 cards or one that should be flipped
- ⑧ **Duck:** 2
- ⑨ **Beaver:** Ok A & 7

Common fallacy: vowel + even number → confirmation bias.

Correct approach: vowel (to check for an even number on the other side) + odd number (to ensure there is not a vowel on the reverse)

Method - Data

Group id	Message id	Message type	Origin	Original text	Card selection
G001	1	INITIAL	SYSTEM	Bee,SYSTEM,Guinea pig,Beaver&&2,7,A,H	{"Beaver": ["7", "H"], "Bee": ["2", "A"], "Guinea pig": ["7"]}
G001	2	MESSAGE	Guinea pig	hi	{"Beaver": ["7", "H"], "Bee": ["2", "A"], "Guinea pig": ["7"]}
G001	3	MESSAGE	Beaver	Hello everyone, so what do you think about the task?	{"Beaver": ["7", "H"], "Bee": ["2", "A"], "Guinea pig": ["7"]}
G001	4	MESSAGE	Beaver	I thought switching 7 and H	{"Beaver": ["7", "H"], "Bee": ["2", "A"], "Guinea pig": ["7"]}
G001	5	MESSAGE	Beaver	would do it, but now after waiting, I am not sure lol	{"Beaver": ["7", "H"], "Bee": ["2", "A"], "Guinea pig": ["7"]}
G001	6	MESSAGE	Bee	So I figured choosing A and 2 would be the correct solution being that vowels have an even number on the other side. This will also be true for choosing an even number card. It will have a vowel on the opposite side.	{"Beaver": ["7", "H"], "Bee": ["2", "A"], "Guinea pig": ["7"]}
G001	7	MESSAGE	Guinea pig	i think h and 2	{"Beaver": ["7", "H"], "Bee": ["2", "A"], "Guinea pig": ["2", "H"]}
G001	8	SUBMIT	Guinea pig	2,H	{"Beaver": ["7", "H"], "Bee": ["2", "A"], "Guinea pig": ["2", "H"]}
G001	9	MESSAGE	Bee	Why H?	{"Beaver": ["7", "H"], "Bee": ["H"], "Guinea pig": ["2", "H"]}
G001	10	MESSAGE	Guinea pig	because vowels not present	{"Beaver": ["7", "H"], "Bee": ["H"], "Guinea pig": ["2", "H"]}
G001	11	SUBMIT	Guinea pig	2	{"Beaver": ["7", "H"], "Bee": ["H"], "Guinea pig": ["2"]}
G001	12	SUBMIT	Bee	2,H	{"Beaver": ["7", "H"], "Bee": ["2", "H"], "Guinea pig": ["2"]}
G001	13	MESSAGE	Beaver	I had at first, but on the other side for me could be not directly next to it. 2A just seems to easy	{"Beaver": ["7", "H"], "Bee": ["2", "H"], "Guinea pig": ["2"]}
G001	14	SUBMIT	Guinea pig	2,A	{"Beaver": ["7", "H"], "Bee": ["2", "H"], "Guinea pig": ["2", "A"]}

Results - Reliability

Validation Phase (Rounds 4 & 5)

Indicators		Round 4 inter-rater reliability		Round 5 inter-rater reliability	
		Agreement percent	Quadratic-weighted Cohen's κ	Agreement percent	Quadratic-weighted Cohen's κ
Human 1 & GPT-5	Affect	98.5%	0.959	100%	1.000
	Attitudes	97.8%	0.964	99.50%	0.988
	Information Sharing	100.0%	1.000	100%	1.000
	Task Mental Model	99.3%	0.990	99%	0.990
	Cross Understanding	98.5%	0.977	97.40%	0.919
Human 2 & GPT-5	Affect	100.0%	1.000	99.50%	0.989
	Attitudes	96.4%	0.952	98.10%	0.972
	Information Sharing	99.3%	0.995	99.30%	0.995
	Task Mental Model	92.0%	0.895	97.90%	0.978
	Cross Understanding	91.2%	0.874	96.70%	0.904
Human 1 & Human 2	Affect	98.5%	0.959	99.50%	0.989
	Attitudes	94.2%	0.915	97.60%	0.960
	Information Sharing	99.3%	0.995	99.30%	0.995
	Task Mental Model	91.2%	0.885	97.40%	0.973
	Cross Understanding	89.8%	0.850	96%	0.890

Results - Reliability

Application Phase (Rounds 6)

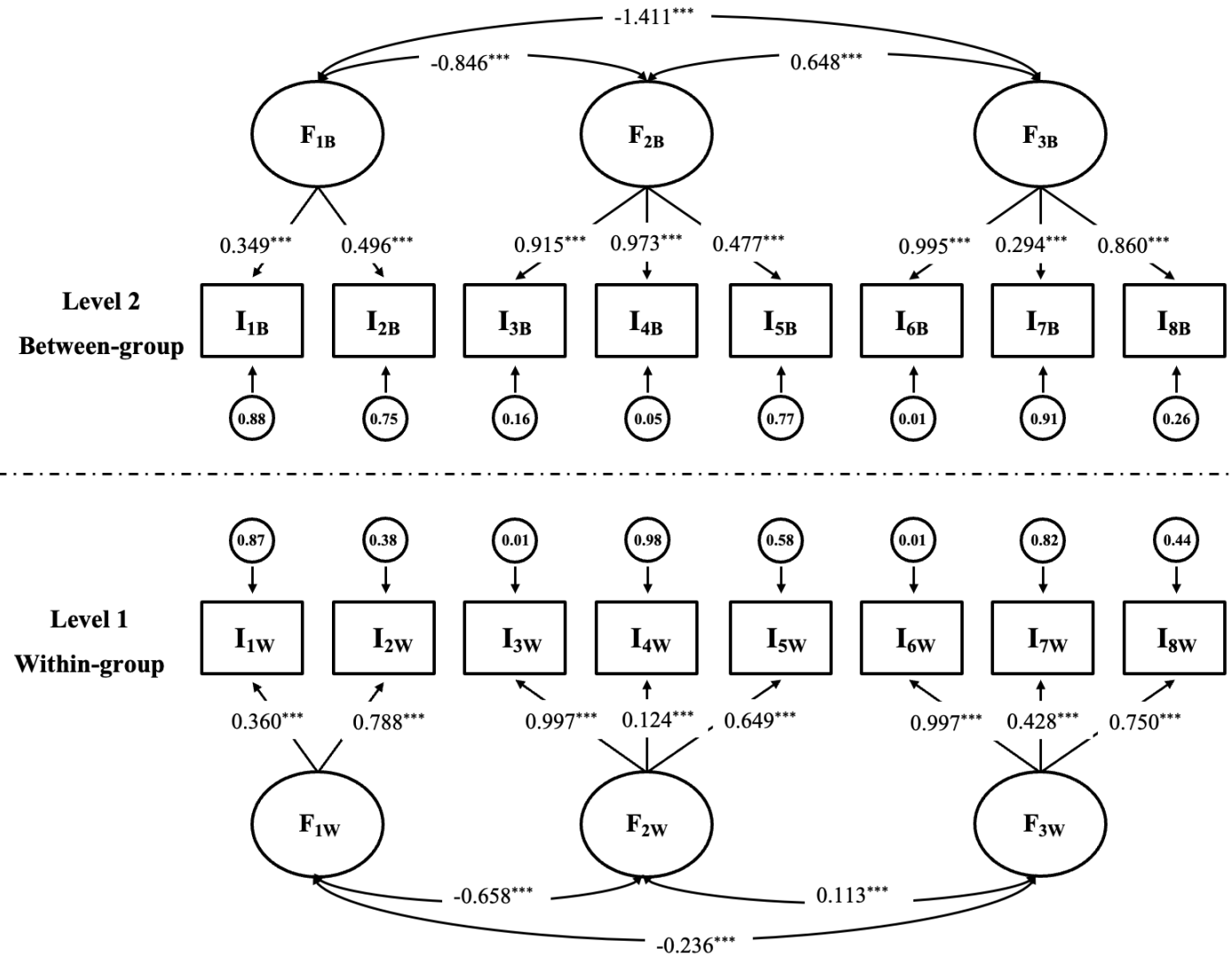
Indicators	Agreement Percent						Quadratic-weighted Cohen's κ					
	GPT5 vs GPT5.5	GPT5 vs Gemini	GPT5 vs Claude	GPT5.5 vs Gemini	GPT5.5 vs Claude	Gemini vs Claude	GPT5 vs GPT5.5	GPT5 vs Gemini	GPT5 vs Claude	GPT5.5 vs Gemini	GPT5.5 vs Claude	Gemini vs Claude
Affect	91.3%	89.1%	89.1%	90.5%	86.8%	86.9%	0.803	0.767	0.752	0.808	0.730	0.735
Attitudes	85.5%	85.8%	84.4%	85.3%	83.8%	86.7%	0.824	0.829	0.798	0.834	0.806	0.835
Information Sharing	81.2%	75.3%	78.6%	80.3%	81.3%	78.8%	0.844	0.761	0.821	0.831	0.850	0.810
Task Mental Model	92.4%	89.2%	89.9%	89.9%	90.7%	89.0%	0.917	0.887	0.894	0.892	0.902	0.887
Cross Understanding	75.3%	68.7%	70.5%	71.5%	71.1%	69.2%	0.664	0.584	0.597	0.608	0.586	0.595

Results - Multilevel SEM

1,272 adult participants in 353 groups
(3-5 person per group)

Three-factor model fit data well, χ^2
(37) = 181.97, $p < .001$, CFI = 0.946,
TLI = 0.918, RMSEA = 0.055, SRMR
within = 0.056, SRMR between = 0.129.

Group-level and individual-level
collaboration skills were computed by
averaging three factor scores.



Results - Qualitative Inspections

Conversation Utterances for Representative Individuals in Three Collaboration Levels

MID	ID	Message	
11	G331_Leopard	ok i agree E and 9.	• Short and simple statement • Less mutual regulation • Reach a rapid consensus
26	G331_Leopard	what about 9 and C. 'i'm not sure, it's confusing'.	
29	G331_Leopard	yea E and 9.	
2	G333_Panda	To test the <u>rule</u> , I chose A to see if there was an even number on the other. side. I prove the rule.	• Explain reasoning explicitly • More task-focused interaction • Emerging awareness of others' perspective
5	G333_Panda	Seems we all are in agreement.	
11	G333_Panda	Well, I'm not sure turning 4 could test the rule. The rule does not state that a cons wouldn't prove the rule to be true.'	
16	G333_Panda	It's the only logical choice that could really test it.'	
1	G061_Emu	hi.	• Sustained and reciprocal interactions • Extensive sense-making process to reach a consensus
4	G061_Emu	I just guessed 6 and E were the answer.	
7	G061_Emu	So <u>we</u> all agree with E?	
10	G061_Emu	But <u>you</u> could verify with G too to see if it reveals a number. even*. 'Isn't there or	
12	G061_Emu	Yeah that's why I don't fully get it haha'. Unless choosing all 4 is the answer.	
18	G061_Emu	okay, we'll try all 4 unless <u>anyone else</u> knows the answer.'. 'Alright I've locked it in.'.	

Note. MID denotes the message ID. G331_Leopard represents a low-level collaboration skill individual, G333_Panda a medium-level, and G061_Emu a high-level.

Future Direction



Stage 1 Feasibility

- Proposing analytical framework

Stage 2 Generalizability

- Real classroom environments
- Diverse group structure
- More theory- and data-driven indicators from video recordings
- Better modelling approach

Stage 3 Implication

- Analytical tools
- Collaboration dashboard
- Real-time feedback generation

Thank you!

Q & A